

CD-ROM AS MEDIUM FOR LARGE GEOGRAPHIC INFORMATION FILES

William W. Wallace, Janet K. Mencher

National Geodetic Survey, NOAA, Silver Spring, MD 20910

Abstract: In June 1993 the National Geodetic Survey chartered a Total Quality Management Process Action Team (PAT) to study the feasibility and practicability of using Compact Disc-Read Only Memory (CD-ROM) as a medium for disseminating large geographic information data sets, specifically geodetic control data and related software. This is a case study of the PAT's progress. Some of the issues discussed include technical issues involving CD-ROM state-of-the-art technology and cost, content issues concerning which data and software are appropriate for this medium, distribution issues addressing the target audience, and management issues concerning personnel and budget. The PAT's study lasted approximately 3 months, and resulted in production of a prototype CD-ROM product which was distributed first to a small group of traditional users of geodetic data for their comments and recommendations. The paper closes with an assessment of the suitability of CD-ROM as a medium for large geographic/geodetic data sets.

INTRODUCTION

Starting in 1992, the National Geodetic Survey (NGS), a division of the National Oceanic and Atmospheric Administration's Coast and Geodetic Survey, began using a Total Quality Management (TQM) approach to planning and problem solving. Under the guidance of an Executive Steering Committee, several Process Action Teams (PATs) were formed during late 1992 and early 1993 to develop various aspects of a strategic plan for NGS. One of the recommendations that resulted was to study the feasibility and practicability of using CD-ROM as a medium for disseminating large data sets containing geographic/geodetic information. Accordingly, a new PAT was formed to conduct this study. The team charter called for the PAT to gather relevant technical information and debate the issues "until substantial agreement is reached." The primary issues defined in the charter were those of data content, technology, management, and dissemination of a product. All of these as well as some additional points are discussed in this paper. The deliverable product for the team, assuming that it found CD-ROM to be a viable option, was a product development plan, describing actions to be taken both to develop the product and at steady state operations.

Because this is a paper on CD-ROM technology, not TQM approaches to management, we will not spend time discussing the details of TQM. Suffice it to say that the team agreed on rules of conduct and operating procedures and followed these throughout the process. We found, as other teams before and after, that the TQM approach works very well when guidelines are agreed to and followed.

CONTENT ISSUES

When team members discussed issues, they were not treated as independent, isolated issues; i.e., content issues were not discussed completely separately from technical issues. However, for ease of presentation in this paper, we have grouped the topics discussed under the categories mentioned. The first category concerns the data and software that would be included on the CD-ROM. Some of the types of questions that we addressed are the following: Should we consider a product which is issued on a regular basis or should we confine ourselves to a data sampler, designed to demonstrate what is possible and develop user interest? What types of geodetic data should be on a CD-ROM? What data elements? What area coverage? What data retrieval software should be supplied? Should this also be on the same CD-ROM? What about "browsing" software? Visualization (graphics) software? Should an instruction (README) file be included on the CD-ROM or should there be separate hardcopy instructions?

Before we address these questions, some description of what is meant by geodetic data is appropriate, since we should not assume that most of our audience is familiar with geodesy. Geodetic data pertain to attributes of geodetic control stations, sometimes called "survey marks" or "bench marks." Over the past 150 years this agency and its predecessors have established over 750,000 of these stations throughout the United States to

provide precise and accurate horizontal (latitude and longitude) and vertical (height or elevation above a common datum) references for engineers, surveyors, and others requiring accurate positions and heights. Collectively these stations make up the National Geodetic Reference System (NGRS) which supplies the framework to support surveying, mapping, charting, navigation, boundary determination, property delineation, infrastructure development, and geographic and land information systems. Examples of attributes that pertain to these stations include stations' names or designations, latitudes, longitudes, heights, state plane coordinate values, datums used, a description of where the mark is located and how to reach it, a history of when the mark was recovered and the condition in which it was found, and a unique identifier that pertains only to that station and ties all station records and attributes together in the NGS data base.

As one can easily imagine, the data files containing all of these station records for over 750,000 marks are quite large (approximately 4 gigabytes or 4 billion characters). Although there have been many types of geodetic data products in the past, the most used (and hopefully useful) product is the NGS "data sheet" which combines the most often requested data elements in a concise format for each station in a given geographic area. Figure 1 is a sample of a two-page data sheet record. Although the data sheet is a concise format, the average station fills the equivalent of nearly two pages, so file sizes even for small geographic areas can still be large. The "standard" medium for disseminating these data sheets for the past several years has been diskette. Often the data are distributed on a county by county basis. On average, one county's data sheet files will fit on one high density diskette, although there are areas which require multiple diskettes for one county. One of the reasons for examining the possibility of CD-ROM (or other high density media) for geodetic data distribution is that many users would like to receive an entire state or several states of data. This can require hundreds of diskettes, which is costly for the user and very time consuming for our data distribution personnel to produce.

We go back now to some of the questions posed on the content issue at the beginning of this section. Should we consider a "sampler" product initially, or a true, useful data product. Our team felt that its members understood the needs and capabilities of its user community sufficiently well that it was comfortable issuing a CD-ROM product following the format of an existing product, rather than exhausting the considerable resources that would be necessary to generate a "new" product to be called a "sampler." Since the digital data sheet is a well-established product with a suite of viewing and extraction tools and existing documentation, the team agreed that the basic content of a regularly issued CD-ROM would be the digital data sheet format. The next question to be answered was "How much data should we put on a CD-ROM?" One state? Several states? The entire country?

The team considered but rejected the possibility of putting the entire U.S. on a single CD-ROM. This would require data set compression, and users would be required to have sufficient disk space to hold decompressed files. In addition, few users would be interested in so much data. Most of NGS' traditional users work in areas covering a few counties, or at most a few states. Since a CD-ROM holds 600 to 680 megabytes(MB) of data, depending upon the type of CD chosen, we set a criterion of 520MB of digital data sheet information as the limit to put on a single CD-ROM. This leaves room for indexes, search and retrieval software, and perhaps some future data sets and/or software that would be desirable to add. Given this limit, the U.S. was partitioned into five regions as shown on Figure 2. The estimated file sizes are shown for each region. Because geodetic control is generally more dense in coastal areas, the CD-ROMs for the Northeast, Southeast, and West regions cover a smaller geographic area than the North and South Central regions.

The remaining content questions concern the types of presentation software and instructions to put on the CD-ROM. The existing digital data sheet product distributed on diskette has presentation software called "DSX" which allows several options for viewing the data. The data file, called "DSDATA," is an ASCII file containing data sheets for one county in alphabetical order. On the CD-ROM there would be an index and multiple file names to reflect the CD-ROM's multiple counties in many states, instead of one county per diskette. Since the data are ASCII files, the CD-ROM should be directly readable on an IBM PC-compatible computer, running DOS, and also on most MacIntosh and UNIX platforms. However, in order to use the DSX presentation software

on a system other than DOS, the source code (which is included on the CD-ROM) for DSX would have to be run through an ANSI C language compiler for whatever platform one is using. The "README" file used on diskettes would have to be modified to reflect the mentioned changes for CD-ROM, but this is a relatively minor task. The team also discussed some possible future additions for the CD-ROM product. These are discussed in the section of this paper called "Product Development Plan."

TECHNICAL ISSUES

Although the team charter called for us to look at CD-ROM as a medium, we also examined other media, both magnetic and optical. Space does not allow us to include all of the details discussed by the team, but highlights are included here.

Magnetic versus Optical Technologies

Examples of magnetic media are floppy and hard disks and magnetic tape. Magnetic media use a magnet in the read/write head to change the polarity of bits on the magnetic media from 0 to 1 or 1 to 0. The head is bidirectional, i.e., it can change bits from 0 to 1 or 1 to 0 in one pass. The primary advantage of magnetic hard disk drives is speed. Hard disks are very fast, with average access times of 20 milliseconds or less. A major disadvantage is that all magnetic media can be affected by stray magnetic fields, thus scrambling of information can occur.

There are three types of optical storage technologies - WORM (Write Once Read Many), rewritable optical, and CD-ROM. Optical disc technology uses a laser beam to read/write to a disc. Because the light from a laser beam is more focused than magnetic read/write heads, more data can be squeezed onto an optical disc in the same relative space. Some general advantages of this technology over magnetic are the following:

They are more durable than magnetic media - most optical discs are made of polycarbonate, the same plastic used in bullet-proof windows.

They have greater storage capacity, are more reliable, removable, and cost far less per megabyte.

Head crashes don't occur because the optical head is held 1000 times farther away from the disc surface than the "flying" magnetic head.

Contamination is not much of a concern since the laser beam is focused on the recording layer below the clear plastic, allowing the drive to read through most scratches and fingerprints.

The primary disadvantage of optical technology is speed. Access time is slower than hard disks due to the weight of the optical head with its laser, lenses, and mirrors. At the present time, access times for optical discs are two to six times slower than high performance disk drives. Because of this, optical drives are not expected to replace hard disks within the next several years.

Media Selection

Among optical technologies, there are many reasons for choosing CD-ROM over WORM or rewritable for our application:

There is a well established, accepted standard for CD-ROM technology known as ISO 9660. The standard defines the media size and characteristics, physical data layout on the disc, error correction, disc rotation speed, and other parameters. This means that every CD-ROM reader which is ISO 9660 compliant can read a CD-ROM. Currently, there are no international standards for WORM technology. WORM discs come in a variety of sizes and densities. There are so many proprietary WORM drives that accessing information from a WORM disc written on a different drive is nearly impossible.

Files can be created on a CD-ROM that are operating system independent. This means that an ISO 9660 disc works identically on a Personal Computer (PC), MacIntosh, or SUN workstation (UNIX).

CD-ROM drives will be backwardly compatible. This means that CD-ROM readers will be able to read a CD-ROM made today on the computers of tomorrow. This backward compatibility is specified in the ISO 9660 standard.

CD-ROMs as a medium are very durable. Only severe damage will make a CD-ROM unreadable. Environmental hazards like magnetic fields, X-rays, moderate heat or extreme humidity, as well as old age, do not hurt CD-ROM data as they do magnetically stored data. CD-ROM discs are also immune to infection by computer viruses.

CD-ROM technology is widely accepted. Although the other optical technologies are interesting, none is as widespread as CD-ROM, and it is not likely that any of them will replace CD-ROM in the marketplace in the foreseeable future.

CD-ROM is appropriate for products intended to have a wide distribution. Once you retrieve and organize data on a CD-ROM glass master, you can replicate the master by stamping out other CD-ROMs from it. This means customers can get the data quickly, easily, and cost effectively. Because you can only write once to a CD-ROM, no one can overwrite data on it. CD-ROMs are also well suited for the distribution of large amounts of data, since a single CD-ROM can hold as much as 680MB of data.

There are several steps involved with making CD-ROMs. The first involves putting data and software in the desired form for the final CD on a medium from which a "glass master" can be made at a manufacturing site. This can be done on magnetic tape or removable hard disk, but these technologies have been outdated by desktop mastering hardware and software which can be purchased for as little as \$5,000 to \$8,000. The recorders used for this are called CD-R recorders, and the process is a chemical one, "burning" the data onto a CD "gold master." CD gold masters are also called "one-offs" because they are not used for mass production. If more than a few copies are needed, the bulk copies are made at a manufacturing site using the gold master to make a glass master from which production copies are made. Making a glass master is a physical process rather than a chemical one, where the data are "stamped" on the disc. Only a few large facilities have the capability of mass producing CD-ROMs.

MANAGEMENT ISSUES

A primary management issue concerns responsibilities. Who is responsible for development of a production capacity? Who is responsible for production once the capacity is established? The Product Development Plan, discussed in a later section, calls for a product manager who has the primary responsibility for executing the Plan. This is the first time in NGS that a product has a designated "manager." The team recommended that one of the authors of this paper, William Wallace, be the product manager. During the development phase, the Systems Development Branch of NGS will support the product manager to develop a productive capacity by executing ADP procurements, assisting with data base retrievals, and developing software. Once a productive capacity has been reached, the National Geodetic Information Center (NGIC) will take over routine parts of the production process and sell and market these products. It is important that NGS understands the impact of this product upon our existing clientele and identifies new clientele as well as their additional needs.

Another management issue concerns the budget for this project. How much funding and how many personnel will be required for development and then steady state operations? Based on discussions held with others in NOAA who have had extensive experience with CD-ROM, notably the National Oceanographic Data Center (NODC) and the National Geophysical Data Center (NGDC), the team learned that costs could be well defined before the fact. The cost of desktop premastering and mastering hardware and software, i.e., for making the gold masters, was described in the previous section to be in the \$5,000 to \$8,000 range at the time of this writing. This has come down considerably in the past couple of years and may continue to do so. The time it takes to do the mastering depends upon several factors such as the size and complexity of the data files, and whether search and retrieval software is to be included. In general, this process will involve one person's time for about 4 to 8 hours, depending upon complexity. The cost of mass producing copies at a manufacturing site will be in the \$2 to \$5 range per copy, depending upon graphics required for the label and jewel case, number of colors needed, etc.

Some other questions for which answers are less well-defined concern decision points. What response from users is expected? What is the minimum response at which this can be considered a viable product? Conversely, under what conditions might it be dropped? Some of the answers to these will be based upon past experience with existing products and NGS' traditional users. This leads us to topics covered in the next section.

DISTRIBUTION ISSUES

These issues are concerned with topics such as the projected market for the CD-ROM product, pricing of the product, and marketing plans and strategies. The NGIC of NGS has over 20 years experience with providing established products to our customers. The current customer mailing list contains approximately 4,500 addresses. Most of these fall in the category of "traditional" users of geodetic data, e. g., surveying and engineering companies and government agencies which require accurate positioning for their daily operations. The size of the typical company on our mailing list is in the range of 3 to 10 employees. Many small organizations are slow to invest in new technologies compared with, for example, a government agency, due to limited resources to purchase state-of-the-art equipment. We saw this with Global Positioning System (GPS) technology during the late 1980s. Although many government agencies and large firms were buying GPS receivers, most smaller companies could not afford the equipment until the prices came down substantially. Such should not be the case with using CD-ROM technology. CD-ROM readers can be purchased for as little as \$200 to \$500. This is a small amount when one considers what may be the cost of not using this technology.

The current data sheet product was described previously as being one county's worth of data on diskette. The price of this is \$30. When we did a preliminary pricing study for the CD-ROM product, we assumed an average production run of 500 CD-ROMs for each region. When production and distribution costs were computed, the cost we determined for the CD-ROM product was \$50. If a customer only needs one county of data, he/she may opt for data sheets on diskette. However, if much more than that is needed, it should be obvious that the CD-ROM product will cost less, even factoring in the cost of a CD-ROM reader. For example, if you want data for the state of Wisconsin on diskette the cost would be \$2,160 (72 counties x \$30). The CD-ROM containing Wisconsin and twelve other states is \$50. Because of this huge price advantage for those who need larger data sets, the team felt that CD-ROM would be welcomed by our traditional users.

How about potential users other than our "traditional" users? The rapid growth and usage of both GPS and Geographic Information Systems (GIS) technologies has led to more interest and use of NGRS data including the data sheet format. Since the NGRS is the most accurate nationwide coordinate system, it serves well as the "base layer" to tie position-related information together in a GIS. Many NGRS stations are also GPS stations and are useful in geodetic applications of GPS. We are trying now to devise marketing strategies which will inform new audiences about the applications of NGRS data and the CD-ROM product.

At the end of the last section, there were questions about decision points. What would be considered a desirable response from users of the CD-ROM product? One suggestion is to see how this compares to the number of requests we receive for the digital data sheets on diskette. If the volume of requests for the CD-ROMs approaches the volume for the established product, we would consider it successful. There will be information about this in the concluding section.

PRODUCT DEVELOPMENT PLAN

This plan was developed because the team believes that new products require a plan that defines the product and assigns responsibilities. The definition is necessary so all parts of the organization can understand the product. The assignment of responsibilities allows different parts of the organization to participate in the creation of the product. It also establishes common expectations about the time line for product development. Product definition and responsibilities have been discussed earlier in this paper. Here we discuss the production schedule, what it will take for a productive capacity, and some auxiliary and possible future products.

The production schedule was determined by trying to realistically estimate steps on the time line. During the period when the team was still meeting, data retrievals were started for the North Central region. This region was picked first because the massive flooding that occurred in the upper Mississippi River Valley during 1993 destroyed considerable geodetic control as well as boundary markers. During the summer and fall of 1993 NGS received more than the average number of requests for data in this region. It took nearly 3 weeks to retrieve data for the 13 states in the region without adversely impacting other computer-related activities at NGS. To be conservative, we would allow at least a month for this for other regions in case there are problems with the retrievals.

Premastering and mastering for the gold master can be done in a day or less if there are no problems. The team has recommended that NGS purchase the necessary equipment to do this. However, since any government procurement takes time, we have made arrangements with NGDC in Boulder, Colorado, to do this until we have our own capability. We decided to FTP (File Transfer Protocol) the data to Boulder over Internet and also send them an exobyte tape as a backup. The mastering equipment at their site is kept extremely busy, so we have to preschedule our mastering several weeks ahead of time. After NGDC has made the gold master successfully, they send it by overnight mail to NGS. Again to be conservative we allow at least a week for this step. Graphics work for the CD-ROM label and jewel case inserts can be done while these first two steps are being completed.

Finally, the last step is sending the gold master and graphics work to a disc manufacturing site for the production of 500 copies of the CD-ROM. Most facilities will turn this around in 5 to 10 working days. Combining these steps comes to approximately 7 weeks. For the five regions to be completed in a yearly cycle, we can stagger their release by 2 to 3 months with some safety factor. The production schedule for the first year's volumes of the CD-ROM is as follows:

<u>Region</u>	<u>Release Date</u>
North Central	10/31/93
Northeast	12/31/93
Southeast	02/28/94
South Central	04/30/94
West	07/31/94

As of March 31, 1994 the first two CD-ROMs have been issued and the third is in production. New issues of the CD-ROMs with the most recent data will be done on a yearly basis on the anniversary of the original release date. Because geodetic data are subject to change, for example when new surveys are added or existing data readjusted, we have to have some provision for updating information contained on a CD-ROM before a new

CD-ROM is made each year. Since CD-ROMs are read-only, updates would have to be done on a separate product. Changes would most likely not affect a large percentage of the data on a CD-ROM, so they may be provided on separate high density diskettes.

To have in-house productive capacity for this CD-ROM product will require purchasing the necessary desktop mastering hardware and software and a dedicated PC with at least a 1 gigabyte hard drive (approximately \$10K to \$12K total). Manufacturing 500 copies at a service bureau will cost about \$2,500 for each region, or about \$12,500 each year. Staffing estimates for data retrievals, graphics, any software needed, and product management are a total of about 0.5 staff years per year.

Future additions to a CD-ROM product may include more sophisticated presentation software with graphics capability. There are several software packages which are being tested which combine in-house written data extraction software with commercially available packages such as CAD. Putting geodetic computation software and/or publications on CD-ROM is another possibility. NGS will canvass its customers to see what types of products are most needed and practicable.

CONCLUSION

The first year's five volume issue of geodetic data sheets on CD-ROM is considered by the PAT to be a prototype product. There will almost certainly be suggestions and recommendations from users to add items or otherwise improve the product. The first volume issued, the North Central region, was initially sent to our state geodetic advisors. There are currently 26 advisors covering 27 states (California and Alaska share an advisor). These advisors work closely with surveyors, engineers, and others within a state to coordinate and help with geodetic activities. Because of this, they are more in touch with users at the local level than most office employees at NGS, and are in a good position to test new products. Their reaction to the CD-ROM product has been almost universally positive. Many of them do not yet have their own CD-ROM reader, but all intend to get one as their state becomes available on CD-ROM.

After the initial positive response from the state advisors, NGS began to actively promote the CD-ROM product to its customers. Sales were slow for the first couple of months until information flyers and advertisements were in circulation. Customer reaction, like the state advisors, was very positive. Most who did not already have the capability to read CD-ROMs said they would purchase readers in the near future. At the time this paper is being written (March 1994), NGS has disseminated nearly 50 percent of its inventory of 1,000 CD-ROMs (500 each from the North Central and Northeast regions). Since most customers are already familiar with the data sheet product on diskette, they adapted easily to the product on CD-ROM. When we question users about what they would like to see added or changed, the most frequent response is some form of viewing software where data could be seen against a map base. Historically, NGS had a paper product called a "geodetic control diagram" which gave this type of presentation on a map base. Since we no longer support this paper product, we had anticipated a high interest in some graphic presentation of our digital products.

The answer to the question "Is CD-ROM a suitable and practical medium for large geodetic data sets?" is a definite "yes!" In fact, we predict that CD-ROM will replace most, but not all, of the demand for digital data on magnetic tape. As CD-ROM readers become more common accessories for PCs, we may also see CD-ROM replace most of the demand for geodetic data on diskette. However, NGS will continue to support all of these media as long as there is a need.

The team agreed that for at least the first two cycles of this product (2 years), NGS should continue to question customers about its suitability and the need for improvement. The willingness of the customer to pay more for an enhanced product should be explored. We will also ask what additional products they would like to see on CD-ROM.

ACKNOWLEDGEMENTS

We would like to recognize other members of the PAT whose work contributed to this effort. They are Dr. Charles R. Schwarz (team leader), Vicki L. Davis, David R. Doyle, John G. Gergen, Dixon Hoyle, John D. Love, and Gary M. Young. We would also like to thank Craig B. Larrimore of NGS, the author of the digital data sheet software, for his help with data retrievals, and David A. Hastings of NGDC for his assistance in making the gold masters for the first issue of the CD-ROM product.